Weekly Report

1 Done

1.1 Paper Revision

I did the following revision.

Section 3:

- Tell which are DRs for security guys (models), which are DRs for data owners (visual designs).
- Cite other studies to support our DRs.
- Split DR3 to seek defenses (corresponding to 4.2) and explain the process (corresponding to visual design).
- Mention that we should provide a recommendation based on data characteristics in DR6 (corresponding to 4.3).

Section 4:

- Highlight DR1.
- Explain why we need to simulate inference attacks and prove the representative of Bayesian network by citing 5 or more paper about inference attacks based on different approaches.)
- Highlight the new DR-seek defenses.
- Highlight DR5.
- Change title to “Scheme Recommendation”.
- Label DRs in figure 2.
- Highlight DRs in text.

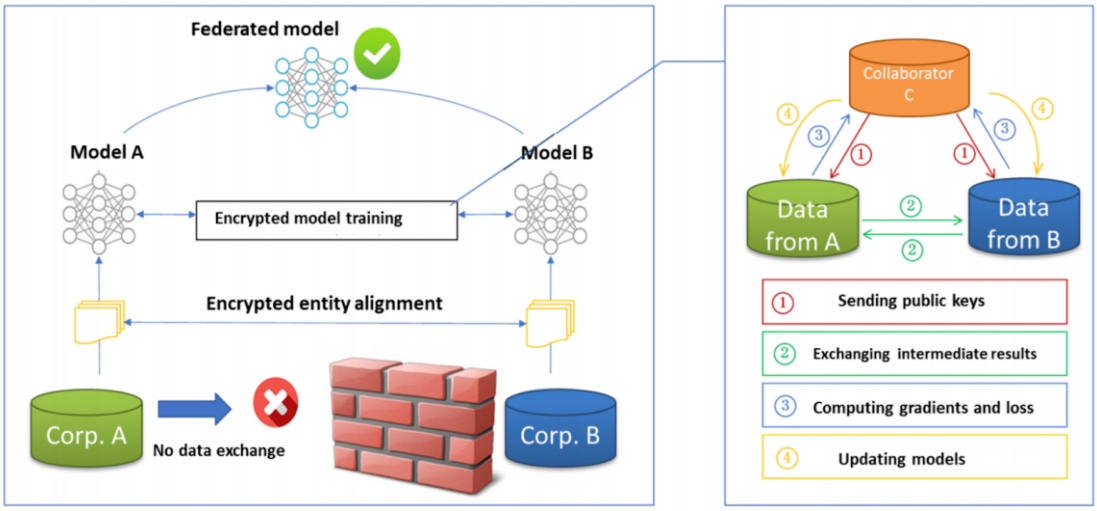
Section 5:

- Add introduction about the graduate dataset, including several sentences about dataset size, why it needs privacy and a table listing relevant graduation data attributes.
- Explicitly reference the DRs.
- Compare with other designs.

1.2 Project Plan

Background:

- Federated learning:



Architecture for a vertical federated-learning system [cite: Federated Machine Learning: Concept and Applications]

Federated learning can help corporations (with a data source) to improve their learning models without the risk of privacy exposure due to no data exchange. However, there is no free lunch. This new approach leads to a new problem:

1. How many benefits can a corporation get from cooperation?
2. How to set the fare for different corporations to participate in such mutually beneficial activities?

Available Datasets:

We need datasets to simulate several data sources with real-time incrementally updated instances.

1. Amazon Review Datasets

http://jmcauley.ucsd.edu/data/amazon/
https://www.kaggle.com/datafiniti/consumer-reviews-of-amazon-products

We can split the dataset by product categories and employ the sub-datasets to simulate reviews from stores selling products in different categories. Reviews or preferences could be sensitive, so they don't want to share datasets with others. But integrating reviews on products in different categories from the same customers could help improve models for prediction or recommendation.

2. My Anime List Dataset

https://www.kaggle.com/azathoth42/myanimelist#anime_cleaned.csv

This dataset contains information about demographic data, anime lists and rating information of users and anime introduction. Those animes can be categorized into TV, OVA, Movie, etc. Similarly, we can suppose they come from different data sources.

Goal:

What I want to show is the difference in the contributions of different datasets to each other.

1. Uncertainty:

   Datasets may have some instances that are hard (or ambiguous) to the learning model. They need a favor of the datasets that can contribute to improving the learning model for these instances. [cite: "*Data Utility Maximization When Leveraging Crowdsensing in Machine Learning*"] Therefore, we need to understand what the corporation has and what he or she can get from others. Actually, the instances mentioned above may also be outliers. Here, we may need some manual judgments.

2. The number of instances included in each iteration:

   The results given by a few records may be sensible to outliers. Thus, how many instances can be included in each iteration is significant to evaluate the data source. But the number of instances may fluctuate by time. We need to show the fluctuation and let human adjust iteration frequency to assure the effectiveness.

3. Response time:

   In each iteration, data source need to update their own models and load new weights and losses to the central server. We need to observe if they have the ability to respond timely. (If we set a long iteration interval, we can be omitted this variant.)

4. Weight differences:

   If the updated weight from a certain source changes dramatically, there may be an abnormality caused by outliers or emergencies. We need to stop it if it is irrelevant to other sources.

5. Loss of different models:

   To verify the effectiveness of the learning model.

Codes:

Federated learning has open source codes that can be found from the following links:

1. TensorFlow:
   https://github.com/tensorflow/federated/blob/v0.4.0/docs/federated_learning.md
2. WeBankFinTech:
   https://github.com/WeBankFinTech/FATE

Both of them don’t provide intermediate results. I may select the first one supported by TensorFlow so that I may have a chance to record intermediate results after modifying their codes.

Design: